NEW ROUTING SCHEMES FOR PACKET SWITCHING NETWORKS

CROSS-REFERENCE TO RELATED APPLICATION

This application is a non-provisional application of provisional application Serial No. 60/212,333 filed June 16, 2000.

BACKGROUND OF THE DISCLOSURE

1. Field of the Invention

The invention relates to methods for efficiently routing a packet in an interconnection network of routers and switches and, more particularly, for providing flexibility and significant complexity reduction in implementation of such networks; and packet switching networks incorporating such methods.

2. Description of the Background Art

A packet switch **100**, sometimes called a "router", is depicted by FIG. 1. Key components in it includes a number of "line interface cards" (or simply called "line cards"), a "switching fabric" **105,** and a routing controller **101**. A physical line card includes two logical parts, the "input module" **103-1** and the "output module" **104-1**, which provide physical interface to packet generators and receivers. The switching fabric **105** provides the infrastructure for moving packets between input modules **103-1, 103-2, 103-3, 103-4** of line cards and output modules **104-1, 104-2, 104-3, 104-4** of line cards. The routing controller **101** uses various routing protocols to exchange information with other packet switches to build and maintain a routing table **102**.

The input module **103-1** of each line card includes or corresponds to a packet-forwarding table **106-1**, which contains partial content of the routing table **102.** Packets, in either fixed or variable lengths, enter the switch via the input module **103-1** of a line card. The input module identifies the routing information from the packet header, searches the packet-forwarding table for a match, and translates into a "switching header". The switching header typically contains the in-band control signals for the switching fabric **105**, QoS (quality of service) information, etc. More often than not in broadband switching, the incoming packet is segmented into "cells" of a fixed length. Assume that this is the case. Before a cell enters the switching fabric, the switching header prefixes the cell for guiding the cell through the switching fabric **105** toward an appropriate output. The cell then arrives at the output module of a line card. The output module reassembles cells back into the packet format.

FIG. 2 depicts the general topology of a "packet switching network" **201** consisting of a number of packet switches. Packet switches in the network are called "nodes" **202-1, 202-2, 202-3, …** , of the network. All nodes are directly or indirectly interconnected to one another through "inter-node links" **203-1, 203-2, 203-3, …** , which are communication channels.

Besides inter-node links, there can be other types of communication channels connected to a packet switch, e.g., channels linking the switching fabric to packet generators/receivers. All these communication channels are hosted by line cards on the packet switch. Those line cards hosting inter-node links **203-1, 203-2, 203-3, …** ,

will be called "inter-node line cards". The remaining line cards hosting **204-1, 204-2, 204-3, ...**, host other communication channels.

When a packet switching network, including the nodes and the inter-node links, is viewed as a large packet switching system, it is functionally equivalent to a large packet switch with a large throughput but enjoys two major advantages over the single-switch equivalent. First, a single packet switch with such a large throughput may not be economically feasible. Second, because of the geographic distribution of the nodes in the network, the average distance from a user to the network is shorter than that to the single packet switch. On the other hand, the deployment of a packet switching network instead of a single packet switch incurs the cost of inter-node line cards. This cost is a substantial overhead because, typically, the bulk of the cost of a packet switch is in the line cards.

A packet enters a packet switching network at its "ingress node and exits from its "egress node". When the ingress node is not the same as the egress node, there may also possibly intermediate nodes on the packet route. Moreover, there may be different routes available between the ingress node and the egress node. A "routing scheme" refers to a method for selecting and setting up a route. The existing variety of routing schemes can be classified into the following two categories. Because of the overwhelming application in IP (Internet protocol) networks, the routing information carried by a packet header will be abbreviated as "the IP address".

Hop-by-hop routing. Each node independently chooses the next hop for an incoming packet. That is, each node analyzes "the IP address", the QoS information, etc. in the packet header and then, based on a routing algorithm/policy or a routing table, chooses a next node for the packet route. In fact, this has been the usual connectionless mode in existing IP networks. The packet-forwarding table in the input module of the line card translates "the IP address" into the output address of the switching-fabric that leads to the next node on the packet route.

Source routing. A packet is forced to follow a particular path through the network, which is set at the ingress node of the packet. Thus the input module of the ingress node translates "the IP address" into a fixed route encoded in the form of a sequence of "next-node identifiers" and affixes the sequence in front of the packet. The input module of every node on the packet route performs protocol processing and peels off the leading next-node identifier in the sequence and uses it as a substitute for "the IP address" in the switching control. Since the next-node identifier is much simpler than "the IP address", the packet-forwarding table for the translation from the next-node identifier is much smaller than that for the translation from "the IP address". Reducing the size of the packet-forwarding table also reduces the space- and time-complexities of the operation. The input module of an inter-node line card needs only perform this reduced operation, and hence its cost can be lower than a regular input module.

As mentioned in the above, there are advantages in a packet switching network over a single large packet switch but there is also the disadvantage in the cost of

inter-node line cards. A source-routing scheme over a packet switching network alleviates this disadvantage by somewhat reducing the cost of the input module of inter-node line cards. However, source coding incurs the overhead in packet formatting by carrying the sequence of next-node identifiers, which encode the whole packet route through the network.

It would be desirable if the cost of inter-node line cards can be further reduced or even eliminated altogether.

SUMMARY OF THE INVENTION

The bulk of the cost of a packet switch is in the line cards. Hence the cost of inter-node line cards is a substantial overhead incurred in the deployment of a packet switching network instead of a single large packet switch. This invention presents two new routing schemes over a packet switching network so that the protocol processing at inter-node line cards is drastically simplified or even eliminated altogether, and similarly for the packet switching networks incorporating such methods. The first scheme is of the source-routing type, and the second is of the hop-by-hop type.

In accordance with a broad method aspect of the present invention, a method for routing a packet having a packet header containing routing information through a packet switching network composed of nodes, includes: (a) in an ingress node of the packet switching network receiving the packet, (i) translating the routing information into a fixed route encoded as a sequence of in-band control signals,

(ii) fragmenting the packet into cells of a fixed length, and (iii) affixing the sequence of in-band control signals in front of each of the cells; (b) deploying each of the in-band control signals in the sequence in a corresponding sequence of nodes on the route to guide each of the cells through the sequence of nodes and consuming said each of the in-band control signals in said corresponding sequence of nodes; and (c) at an egress node of the packet switching network, reassembling the cells into the packet.

BRIEF DESCRIPTION OF THE DRAWINGS

The teachings of the present invention can be readily understood by considering the following detailed description in conjunction with the accompanying drawings, in which:

FIG. 1 is a block diagram of a packet switch of the prior art;

FIG. 2 is a block diagram of a packet switching network of packet switches;

FIGS. 3A-C depict the cell formats in the inventive source-routing scheme.

FIG. 4 is a flow diagram of the inventive source-routing scheme;

FIG. 5 is a block diagram of multiplexing, transmission, and demultiplexing of data through an inter-node link;

FIG. 6 is a block diagram of a switching fabric pertaining to the concept of output grouping;

FIGS. 7A-C depict the cell formats in the modified version of the inventive source-routing scheme;

FIGS. 8A-C depict the cell formats in the inventive hop-by-hop routing scheme; and

FIG. 9 is a flow diagram of the new hop-by-hop routing scheme.

To facilitate understanding, identical reference numerals have been used, where possible, to designate identical elements that are common to the figures.

DETAILED DESCRIPTION

The present invention improves on the conventional source-routing scheme stated in Description of the Background Art. Certain commonly seen characteristics of the packet switching network are assumed:

(1) Packets are routed through the network in the form of fixed-length cells.

(2) There are at most N nodes on the route of a packet through the network. For example, N can be as small as 2 only when there exists an inter-node link from every node to every other.

In principle QoS information can be included in the switching header of the cells as part of the in-band control signal. However, in order to simplify the descriptions of the embodiment, the inclusion of the QoS information will be omitted.

1. One illustrative embodiment – a new source-routing scheme

Under the conventional scheme of source routing, the input module of the ingress node translates "the IP address" into a fixed route encoded in the form of a sequence of "next-node identifiers". The input module of every node on the packet route peels off the leading next-node identifier in the sequence and uses it in the switching control as a substitute for "the IP address". One embodiment of the present invention makes the following highlighted modifications over conventional source routing:

(1) The code for the fixed route is affixed as a switching header in front of every cell of the packet instead of the packet itself.

(2) The input module of only the ingress node fragments the packet into cells, and the output module of only the egress node reassembles cells into the packet. In between, the switching header of every cell carries the routing information to guide the cell through the switching fabric of every node on the route.

(3) The code for the fixed route is composed of a sequence of "in-band control signals" instead of next-node identifiers. Switching control at every node on the route consumes the leading in-band control signal in the sequence. At each node, the in-band control signal, without the need of further translation, guides the cell through the switching fabric. Except for the case of the egress node, the cell is switched into one of the outputs designated for feeding into an outgoing inter-node link; the identity of this inter-node link implies the identity of the next node on the route.

(4) A bi-directional inter-node link may be regarded as a pair of uni-directional inter-node links in opposite directions. The output module of the line card at the originating end of a uni-directional inter-node link processes the cell for the transmission, and the input module of the line card at the terminating end of that link processes the reception. Besides such processing for the purpose of transmission/reception per se, there is no protocol processing at inter-node line cards. Thus the switching fabrics at all nodes are logically integrated into a single in-band-control switching fabric and the inter-node links become logically equivalent to interconnection lines among elements in the single switching fabric.

Assume that there are a total of n nodes, $1 \leq n \leq N$, on the fixed route of a particular packet, including the ingress node and the egress node. The input module of the line card at the ingress node, which is not an inter-node line card, uses the packet-forwarding table to translate "the IP address" of a packet into a fixed route encoded in the form of a sequence of n in-band control signals, each for the switching control over the switching fabric of each node on the packet route. Meanwhile, the packet is fragmented into cells. The sequence of in-band control signals is then affixed in front of every cell. The cell format is illustrated in FIG. 3A before the cell enters the switching fabric of the ingress node. The symbols S_1 **301-1**, S_2 **301-2**, … , S_n **301-3** stand for the in-band control signals to guide the cell through the switching fabric in the n sequential nodes. When $n<N$, a space filler **310** of N–n times the length an in-band control signal is appended at the end. Switching control at the ingress node consumes the in-band control signal control S_1 **301-1**. Afterwards, the cell format becomes as depicted by FIG. 3B if $n>1$, that is, if the ingress node is not the egress node; the cell in this format is then transmitted through the inter-node link without protocol processing at either end of the link (except for the processing for the purpose of transmission/reception per se) and enters the switching fabric of the next node. Switching control at the second node on the route, if $n>1$, consumes the in-band control signal control S_2 **301-2**. And so on. FIG. 3C depicts the cell format upon entering the egress node. Space fillers **303** and **304** are increased to account for the consumed signal controls.

FIG. 4 presents the flow diagram of this new source-routing scheme, as follows:

Processing block **405**: translate "the IP address" into a sequence of in-band control signals

Processing block **410**: segment the packet into cells

Processing block **415**: affix in-band control signals in front of every cell

Processing block **420**: switching fabric switches the cell using the leading control signal in the sequence

Processing block **425**: determine if the node is an egress node

Processing block **430**: if not, go through inter-node link

Processing block **435**: if so, reassemble cells back into the packet format in the output module of egress node

This source-routing scheme in accordance with the present invention eliminates protocol processing at both ends of an inter-node link except the processing for the purpose of transmission/reception per se. Thus the inter-node line cards are can be regarded as eliminated or reduced to simple transceiver cards. Switching fabrics at all nodes and the inter-node links are logically integrated into a single in-band-control switching fabric, and every inter-node link is rendered an interconnection line in this logical in-band-control switching fabric.

2. Elimination of serial-parallel conversion by the new source-routing scheme

In fact, the new source-routing scheme sometimes also simplifies the processing for transmission/reception at the two ends of an inter-node link. In a broadband communication network, inter-node and inter-network transmission is often over a medium at an ultra-broad bandwidth (e.g., gigabits per second), such as an optic fiber or even a wavelength over a wavelength-division optic fiber. On the other hand, the typical transmission bandwidth over an electrical wire in is typically only megabits per second. Thus inter-node transmission in packet switching very often can be described as follows and depicted by FIG. 5. At the output module of an inter-node line card, transmission over, say, K wires **501-1, 501-2, … , 501-3** are multiplexed by the multiplexer **502** into a single-stream transmission **503** over the inter-node link. At the input module of an inter-node line card, the single-stream transmission **503** is demultiplexed by the demultiplexer **504** into transmission over the K wires **505-1, 505-2, … , 505-3**. They are two ways to interpret this system of inter-node transmission:

Most common interpretation. The K wires **501-1, 501-2, … , 501-3** represent a K-wire parallel bus. The natural form of a packet upon its generation is in form of a serial bit stream, but this interpretation assumes that somehow the packets have been converted into the K-bit parallel form before getting on the K-wire parallel bus. The multiplexer **502** gathers bits from the K wires on a rotational basis. Over the single-stream transmission **503** the packets are in the serial-bit form. The K wires **505-1, 505-2, … , 505-3** again represent a K-wire parallel bus. The demultiplexer **504** distributes bits into the K wires on a rotational basis and thus the packets become in the K-bit parallel form

again on the K-wire parallel bus. After going through the inter-node transmission and before arriving at the ultimate receiver of the packets, there has to be the conversion back into the natural form of serial bit stream.

<u>Alternative interpretation.</u> The K wires **501-1, 501-2, … , 501-3** are K separate serial-bit transmission wires, so are the K wires **505-1, 505-2, … , 505-3**. The multiplexer **502** gathers bits from the K wires on a rotational basis. Over the single-stream transmission **503** bits from the K wires are rotationally interleaved. The demultiplexer **504** distributes bits into the K wires on a rotational basis and thus the packets on each of the wires **505-1, 505-2, … , 505-3** are back into the natural form of serial bits.

Now suppose that a packet routed through the switching fabric at a generic node is in the form of serial bits. Let the new source-routing scheme adopt inter-node transmission as depicted by FIG. 5 under the above alternative interpretation. Then, there is no need of conversion between the serial- and parallel-bit forms throughout the whole packet switching network. In other words, the cost of the serial-to-parallel and parallel-to serial conversions is saved from somewhere in the network.

<u>3. Modification on the new source-routing scheme via output grouping</u>

The overhead in cell formatting that the new source-routing scheme incurs is equal to N in-band control signals, where N is the maximum number of nodes on a packet route. There is always some practical limitation on the overhead in cell formatting.

This motivates the search of ways to reduce the overhead. Toward this goal, this section modifies the new source-routing scheme when the switching fabric in a generic node of the network pertains to the concept of output grouping.

Consider an in-band-control switching fabric with the following extra characteristics:

(F1) Some or all of the outputs of the switching fabric at a generic node belong to "output groups". A packet may be destined for an output group instead of a single output. In that case, all members within the output group are exchangeable and the objective is to route the packet to any member in the group.

(F2) The in-band control signal that guides a packet toward an output group is in a fixed length shorter than that for guiding the packet toward a single output.

(F3) If a member of an output group feeds into the output module of an inter-node line card, then so do all other members of the group.

Example. An exemplifying in-band-control switching fabric with these characteristics is "2-stage switching fabric" depicted by FIG. 6. At the first stage is a switching fabric **601** with all its outputs belonging to output groups **611, 612, … , 613, 614** of size K. There are two types of output groups:

Type 1: An output group that feeds directly into the output module of a line card, as exemplified by the output group **612**.

Type 2: An output group that feeds through a K×K second-stage switching fabric into the output module of a line card, as exemplified by the output group **611** that feeds into the second-stage switching fabric **602-1**.

A cell entering the 2-stage switching fabric is destined for either a Type-1 output group or a single output of a second-stage switching fabric. In the latter case, the in-band control signal comprises two parts, the first part for guiding the cell to the proper output group and the second part for further guiding the cell through the second-stage switching fabric to a particular output. On the other hand, the in-band control signal in the former case is just the first part.

Let the switching fabric at a generic node of the packet switching network be subject to (F1) through (F3). When a cell enters the switching fabric of a non-egress node, it is destined for the output group that feeds into the output module of an inter-node line card. Thus the in-band control signal for guiding the cell through this fabric is of the shorter fixed length. The new source-routing scheme described in Section 1 can then be modified so that all but the final in-band control signals to be affixed in front of every cell are in the shorter fixed length. The cell format before the cell enters the switching fabric of the ingress node is illustrated in FIG. 7A, where n is the number of nodes on a particular route. The n-1 in-band control signals s_1 **701-1,** s_2 **701-2,** … , s_{n-1} **701-3** are short ones, while the final in-band control signal S_n **301-3** is of the normal length. When $n<N$ (N is the maximum number of nodes on a route), a space filler **710** (similar to **310)** of N–n times the length a short in-band control signal is appended at the end.

Switching control at the ingress node consumes the in-band control signal control s_1 **701-1**. Afterwards, the cell format becomes as depicted by FIG. 7B; if n>1, that is, if the ingress node is not the egress node, the cell in this format is then transmitted through the inter-node link without protocol processing at either end of the link (except the processing for the purpose of transmission/reception per se) and enters the switching fabric of the next node. Switching control at the second node on the route consumes the in-band control signal control s_2 **701-2**. Space filler node **703** serves the same purpose as filler **303**. And so on. FIG. 7C depicts the cell format upon entering the egress node. The length of the space filler **704** in this format is N-1 times the length of a short in-band control signal. In contrast, the length of the space filler **304** in the cell format in FIG. 3C is N-1 times the normal length an in-band control signal.

4. An illustrative hop-by-hop routing scheme

The new source-routing scheme incurs an overhead in cell formatting equal to N in-band control signals, where N is the maximum number of nodes on a packet route. The modified version in Section 3 again incurs an overhead proportional to N. Such overhead can be infeasible for some packet switching networks, especially those networks with large values in N. This section adapts the new source-routing scheme of Section 1 into a new hop-by-hop routing scheme, where the overhead in cell formatting is independent of N.

The input module of the line card at the ingress node, which is not an inter-node line card, uses the packet-forwarding table to translate “the IP address” of a

packet into a "route tag" and the in-band control signals for the switching control over the switching fabrics of both the ingress node and the egress node (i.e., the destination node). Meanwhile, the packet is fragmented into cells. The sequence of in-band control signals is then affixed in front of every cell. The cell format is illustrated in FIG. 8A before the cell enters the switching fabric of the ingress node. The symbols S **801-1** and T **801-3**, respectively, stand for the in-band control signals over the switching fabrics of the ingress node and the egress node. The symbol R **802-1** stands for the route tag, which is for the use of the next node on the route. One particular design of the route tag is simply an identifier of the egress node.

In the degenerated case when the ingress node is also the egress node, "the IP address" is translated into just the in-band control signal over the switching fabric of the egress node, and the cell format is as depicted by FIG. 8C.

Switching control at the ingress node consumes the leading in-band control signal, which is S **801-1** when n>1 and is T **801-3** when n=1. Assume that n>1. After the switching fabric of the ingress node, the cell format becomes as depicted by FIG. 8B. The cell in this format is then transmitted through the output module of an inter-node line card without protocol processing (except for the processing for the purpose of transmission per se) and arrives at the input module of an inter-node line card of the next node. The input module of an inter-node line card maintains a packet-forwarding table, which determines whether a route tag indicates the local node is the egress node of the

cell (This table is relatively small, since the route tag is much simpler than “the IP address”.)

If so, the route tag is stripped so that the cell format becomes as in FIG. 8C.

Else, the small packet-forwarding table also maps the route tag to the in-band control signal for guiding the cell through the local switching fabric and a new route tag (for the use by the subsequent node on the route). The cell format before entering the switching fabric becomes as in FIG. 8A.

The cell then enters the local switching fabric. In both cases, the cell format is identical to that when the cell was entering the switching fabric of the ingress node. Thus the same procedure can be reiterated at each subsequent node. Finally, the cell reaches the output module of a line card at the egress node, which is not an inter-node line card. The output module receives cells of a packet and reassembles them back into the packet format.

FIG. 9 presents the flow diagram of this new hop-by-hop routing scheme, as follows:

Processing block **905**: translate "the IP address" into a route tag and the in-band control signals for the switching control of both the ingress node and the egress node

Processing block **910**: segment the packet into cells

Processing block **915**: affix in-band control signals in front of every cell

Processing block **920**: switching fabric switches the cell using the control signal for the ingress node

Processing block **925**: go through inter-node link

Processing block **930**: determine if an egress node (checked in the input module of line card)

Processing block **935**: if not, translate the route tag into in-band control signal for the switching control over the switching fabric and a new route tag

Processing block **940**: switching fabric switches the cell using the leading control signal

Processing block **945**: go through inter-node link

Processing block **950**: if so, switching fabric switches the cell using the control signal for the egress node

Processing block **955**: reassemble cells back into the packet format in the output module of egress node.

Although various embodiments which incorporate the teachings of the present invention have been shown and described in detail herein, those skilled in the art can readily devise many other varied embodiments that still incorporate these teachings.